iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.68575

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Data Download Duplication Alert System

P. Praveen[1], P. Lakshmi Sowmya[2], T. Mounisha[3], P. Purandhar[4], Dr. J. N. Swaminathan[5]
[1, 2, 3, 4]*Artificial Intelligence and Data Science JNN Institute of Engineering*
[5]*Professor, J.N.N Institute of Engineering*

Abstract: ***In today's digital world, data download duplication is a huge challenge for organizations dealing with large digital assets. The solution comes in the form of a "Data Download Duplication Alert System" that not only identifies and prevents duplicate downloads but also manages them in real-time. It is vital to reducing redundant storage of data, minimizing bandwidth usage, and improving operational efficiency. This abstract presents the idea, design, and advantages of using a system for detecting and informing users of duplicate data downloads. The main purpose of a Data Download Duplication Alert System is to provide effective resource utilization without compromising data integrity. The system operates by scanning download requests in real-time, matching file metadata like name, size, hash values, and timestamps with records of previously downloaded files kept in a central database. When a match is found, the system triggers alerts, notifying the user or administrator of the duplication. Sophisticated implementations can also provide for user-specified actions, including blocking the duplicate download or allowing it under certain conditions.***

The system architecture in this case is generally composed of a metadata central database for storing metadata, a hashing of files for identification of uniqueness, and a real-time monitoring feature integrated within the downloading process. Machine learning can be used to detect duplication patterns, forecast future duplicate activity, and also improve system performance. The system can be made compatible with organizational policies to generate personalized responses for duplication events such that the working needs of various industries are fulfilled

Keywords: de-duplication, cloud computing

I. INTRODUCTION

Data download has today become a permanent component of the day-to-day lives of both individual and institutional stakeholders. Due to the consistently large amount of data being consumed, shared, and downloaded, a problem to check and limit repeated downloads arose, which becomes very important at times. The challenge is rectified by an Implementation of Data Download Duplication Alert System as a strong technique of identifying redundant downloading, signaling them, and banning them from continuing. This system not only maximizes the use of resources but also helps to improve operational efficiency, minimize costs, and provide improved user experience.

The phenomenon of download duplication is when the same data files are downloaded over and over again, either deliberately or accidentally. This can occur because of network disruptions, unawareness of previous downloads, or the lack of a central system to track and control downloads. Although one duplicate instance might not be significant, the overall effect can be overwhelming, especially for systems with intensive data traffic or constrained storage. Duplicate downloads create unnecessary bandwidth use, redundant use of storage space, and delay in accessing critical resources for companies. These effects can also disrupt business processes, raise operational costs, and contribute to the general inefficiency of digital systems.

A Data Download Duplication Alert System is created to address these issues effectively. Fundamentally, the system works by tracking download requests in real-time and matching file metadata with a database of past downloaded files. Metadata fields like file name, size, type, and cryptographic hash values (e.g., MD5 or SHA-256) are used to identify files uniquely. Once a duplication is identified, the system immediately notifies the user or system administrator to facilitate decision-making. Users can be notified through multiple communication channels like email, pop-up messages, or dashboard notifications. Additionally, the system can be set up to block or allow downloads based on established rules and organizational policies.

Its application entails a range of elements that encompass a metadata repository, which is centralized; hashing algorithms on files; and a real-time monitoring framework. Advanced systems further use machine learning algorithms to foretell duplication trends and enhance accuracy of detection. All these work together to guarantee the system's scalability, dependability, and versatility to fit varied organizational needs. Integration with the current IT infrastructure, including download managers, content delivery networks (CDNs), and cloud storage services, further increases the usability and effectiveness of the system.

©IJRASET: All Rights are Reserved

The advantages of a Data Download Duplication Alert System are two-pronged. It maximizes bandwidth utilization by eliminating duplicate downloads, thus lessening the load on network resources. It also reduces storage needs by eliminating duplication, which is especially important for organizations that deal with large datasets. Also, by making downloads more streamlined, the system improves user productivity and satisfaction. In industries that have stringent data governance regulations, e.g., finance, education, or healthcare, the system promotes compliance through ensuring data traceability and integrity.

II. LITERATURE SURVEY

1) SAM: A Semantic-Aware Multi-tiered Source De-duplication Framework for Cloud Backup

Current de-duplication solutions available in cloud backup scenario either achieve high compression ratio with a heavy de-duplication overhead in terms of higher latency and lower throughput, or have minimal de-duplication overhead with the price of low compression ratio leading to very high data transmission cost, thus creating very high backup window. Here, we introduce SAM, a Semantic-Aware Multitiered source de-duplication system which initially integrates the global file-level de-duplication and local chunk-level deduplication, and additionally utilizes file semantics at each phase in the system, to achieve an optimal balance between deduplication efficiency and de-duplication overhead and eventually reduce the backup window compared to previous methods. Our experimental results with real-world datasets demonstrate that SAM not only provides a better de-duplication efficiency/overhead ratio compared to current solutions, but also reduces the backup window by 38.7% on average.

2) CAB dedupe: A Causality-Based Deduplication Performance Booster for Cloud Backup Services

Since WAN (Wide Area Network), which facilitates cloud backup services, has relatively low bandwidth, the backup time as well as the restore time within the cloud backup infrastructure are desperately in need of minimization in order to facilitate cloud backup as a viable and economical service for telecommuters and small business enterprises alike. Some existing solutions using the deduplication technology to offer cloud backup services concentrate only on excluding repeated data from being transferred during backup processes in an effort to limit backup time, and do not take into consideration much the restore time which we believe is a critical parameter and influences the end-to-end quality of service of the cloud backup services. Herein, we introduce a Causality Based deduplication performance enhancer for both cloud backup and restore operations, named CAB dedupe, which records the causal relationship between chronological versions of datasets processed in consecutive backups/restores, to eliminate the unchanged data from transmission during not only backup operations but also restore operations, thereby to enhance both the backup and restore performances. CAB dedupe is an orthogonal middleware that can be integrated into any current backup system. Our rigorous experimentation, where we integrate CAB dedupe into two current backup systems and supply real world datasets, demonstrates that the backup time and the restore time are both substantially improved, with the reduction ratio reaching up to 103:1.

3) SHHC: A Scalable Hybrid Hash Cluster for Cloud Backup Services in Data Centers

Data deduplication mechanisms are perfect remedies to minimize bandwidth and storage capacity demands for cloud backup solutions within data centers. Existing data deduplication implementations are based on matching fingerprints (hash values) of data segments to detect redundant data and cache the fingerprints in a centralized server. This process has the constraint of limiting total throughput and concurrency performance in high scale systems. Also, the sluggish seek time of hard disks worsens the performance of hash lookup operations that are predominantly random I/Os. In this work we introduce a scalable hybrid hash cluster (SHHC) to support a low-latency distributed hash table for data fingerprint storage. Each cluster's hybrid node consists of RAM and Solid State Drives (SSD) to leverage the high-speed random access inherent in SSDs. This distributed model makes the system scalable, evenly distributes the load on the hash store and reduces the latency of the hash lookup operation considerably.

A. Existing System

The current mechanisms for handling data downloads tend to be based on simple download managers or network monitoring software. Such software is mostly geared towards offering features such as pausing, resuming, and scheduling downloads, but not duplication handling. Duplicate detection is handled manually by users in certain instances, who have to identify and remove duplicate files themselves. Organizations might also depend on conventional file management systems that provide few features, such as comparing file names to identify duplicates. These systems do not have a centralized, automated, and real-time facility to thoroughly identify and manage duplicate downloads. Also, these tools may not integrate within larger IT infrastructures or issue proactive notifications, rendering them ineffective in high-demanding scenarios.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

B. Disadvantages

1) Manual Intervention Required: Existing systems rely heavily on users to manually identify and manage duplicate downloads, which is time-consuming and prone to errors.

2) Lack of Real-Time Detection: Duplicate detection often occurs after downloads are completed, leading to unnecessary bandwidth and storage usage before duplicates are identified.

3) No Proactive Alerts: The users are not alerted about duplication while downloading, and thus the system cannot block duplicate downloads in real-time**.**

C. Proposed System

The suggested Data Download Duplication Alert System is an extensive solution that will overcome the limitation of current systems by offering real-time detection, alerting, and management of duplicate downloads. It works through tracking download requests, examining file metadata like names, sizes, timestamps, and hash values, and comparing them with a central repository of previously downloaded files. When a duplicate is detected, the system automatically notifies the user or administrator and offers configurable features to block, permit, or control the download according to pre-defined policies. Features such as machine learning algorithms can be incorporated to identify patterns of duplicate downloads and make detection more accurate. The platform natively integrates with existing IT architectures, including cloud storage platforms, content delivery networks (CDNs), and organizational download managers, rendering it scalable and flexible across various environments.

D. Advantages

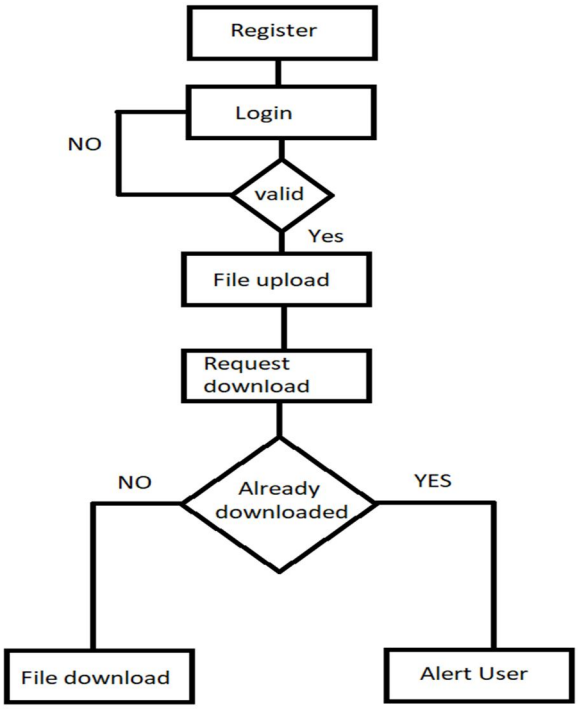
Real-Time Detection and Alerts

Identifies duplicate downloads in real time and notifies users prior to the completion of the download, facilitating proactive handling.

Resource Optimization

Conserves bandwidth, storage space, and processing power by avoiding unnecessary duplicate downloads.

III. METHODOLOGY



IV. BACKEND INTEGRATION MODULE

A. Registration

New User Registration is a framework that enables a user to sign in to the Campus Solutions system in order to complete a specific file sharing. Going through New User Registration, the user can either create a user ID and password, or use an existing user ID to Login to your system. The Login Module is a portal module that allows users to type a user name and password to log in.

B. Login

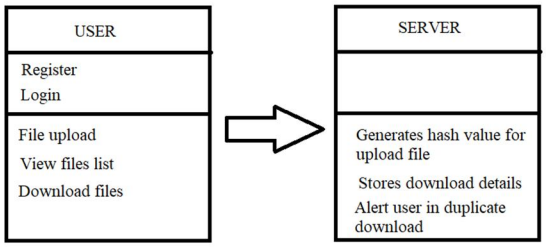
Logging in is typically employed to access a certain page, which intruders cannot access. After the user has logged in, the login token can be employed to monitor what actions the user has performed while online to the site.

C. Upload Files

User select a file to upload it, User can't select the file which has already been uploaded. File validation is done on the basis of the file content not on the filename, So same filename can be used again if the content is different

D. Download Files

In this module user is able to download the file uploaded by some other user. User can download the file just once. If the user attempted to download the file again then user will be notified that the file has been already downloaded by yourself and the downloaed information will be forwarded to the user through mail.

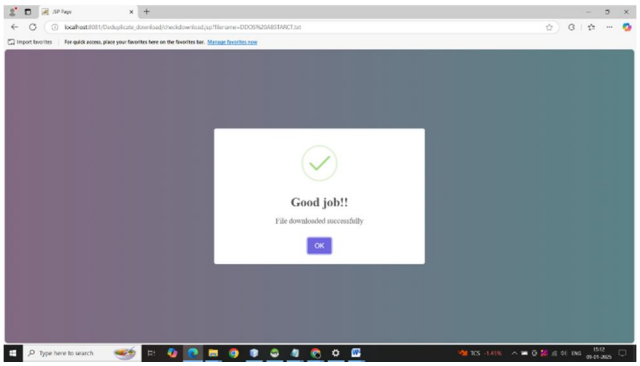


V. NETBEANS

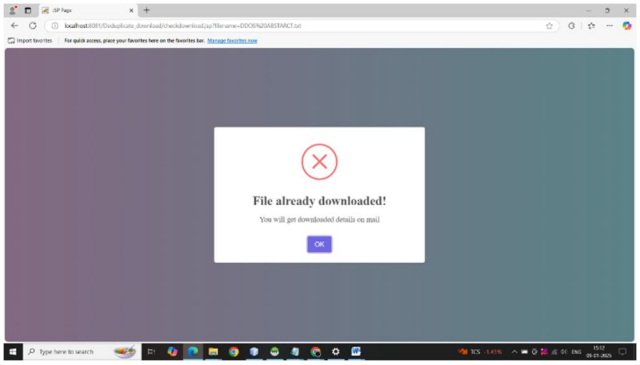
We employed the application NetBeans for backend development.it is solely concerned with the specialization of java programming. Just like application servers, e.g., GlassFish or WebLogic, provide lifecycle services to web applications, the NetBeans runtime container provide lifecycle services to Java desktop applications. Application servers know how to assemble web modules, EJB modules, and associated artifacts, into a single web application. The modularity of a NetBeans Platform application provides you with the ability to satisfy sophisticated requirements by integrating multiple small, simple, and easy-to-test modules that encapsulate coarsely-grained application features

A. System Output

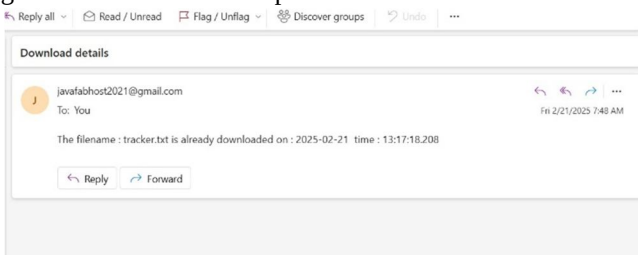
After the successful downloading of file into the data base. It shows the pop-up message that says the file downloaded successfully.



If the file is already in the data base it warns us that the file already exits. The below figure shows us the warning message of that file.



The next figure shows the mail that gives alert about the deduplication in the data base.



VI. CONCLUSION

With the world more dependent on digital assets, handling data downloads in an efficient manner is now an absolute requirement. Data Download Duplication Alert System is a proactive and innovative method of addressing the age-old problem of duplicate downloads. Through the utilization of cutting-edge technologies like real-time monitoring, metadata analysis, and machine learning, the system helps ensure that organizations and individuals can streamline their digital infrastructure without wasting resources. In contrast to current systems, which tend to be resource-greedy and reactive, the solution presented here offers real-time warnings and configurable policies that allow users to avoid duplicate downloads prior to their occurrence.

The capacity of this system to converge without difficulties in a variety of IT infrastructures and conform to changing organizational demands makes it very scalable and future-ready. Apart from functional advantages like bandwidth and storage optimization, the system also strengthens data integrity, compliance, and end-user satisfaction. Its predictive analytics functions further empower organizations to foretell and prepare for prospective duplication scenarios, lending a strategic advantage.

REFERENCES

[1] Y. Tan, H. Jiang, D. Feng, L. Tian, Z. Yan and G. Zhou, "SAM: A Semantic-Aware Multi-tiered Source Deduplication Framework for Cloud Backup", 2010 39th International Conference on Parallel Processing, pp. 614-623, 2010.

[2] W. Leesakul, P. Townend and J. Xu, "Dynamic Data Deduplication in Cloud Storage", 2014 IEEE 8th International Symposium on Service Oriented System Engineering, pp. 320-325, 2014.

[3] Y. Tan, H. Jiang, D. Feng, L. Tian and Z. Yan, "CABdedupe: A Causality-Based De-duplication Performance Booster for Cloud Backup Services", 2011 IEEE International Parallel Distributed Processing Symposium, pp. 1266-1277, 2011.

[4] L. Xu, J. Hu, S. Mkandawire and H. Jiang, "SHHC: A Scalable Hybrid Hash Cluster for Cloud Backup Services in Data Centers", 2011 31st International Conference on Distributed Computing Systems Workshops, pp. 61-65, 2011.

[5] D. Harnik, B. Pinkas and A. Shulman-Peleg, "Side Channels in Cloud Services: De-duplication in Cloud Storage", IEEE Secur. Priv, vol. 8, no. 6, pp. 40-47, Nov. 2010

[6] Walid Mohamed Aly, Hany AtefKelleny, "Adaptation of Cuckoo Search for Documents Clustering," International Journal of Computer Applications (0975 - 8887), Volume 86 - No 1,2014.

[7] Min Li, Shravan Gaonkar, Ali R. Butt, Deepak Kenchammana, an Kaladhar Voruganti, "Cooperative Storage-Level Deduplication for 110Reduction in Virtualized Data Centers," IEEE International Symposiumon Modeling, Analysis & Simulation of Computer and Telecommunication Systems,pp.209-218, 2012.

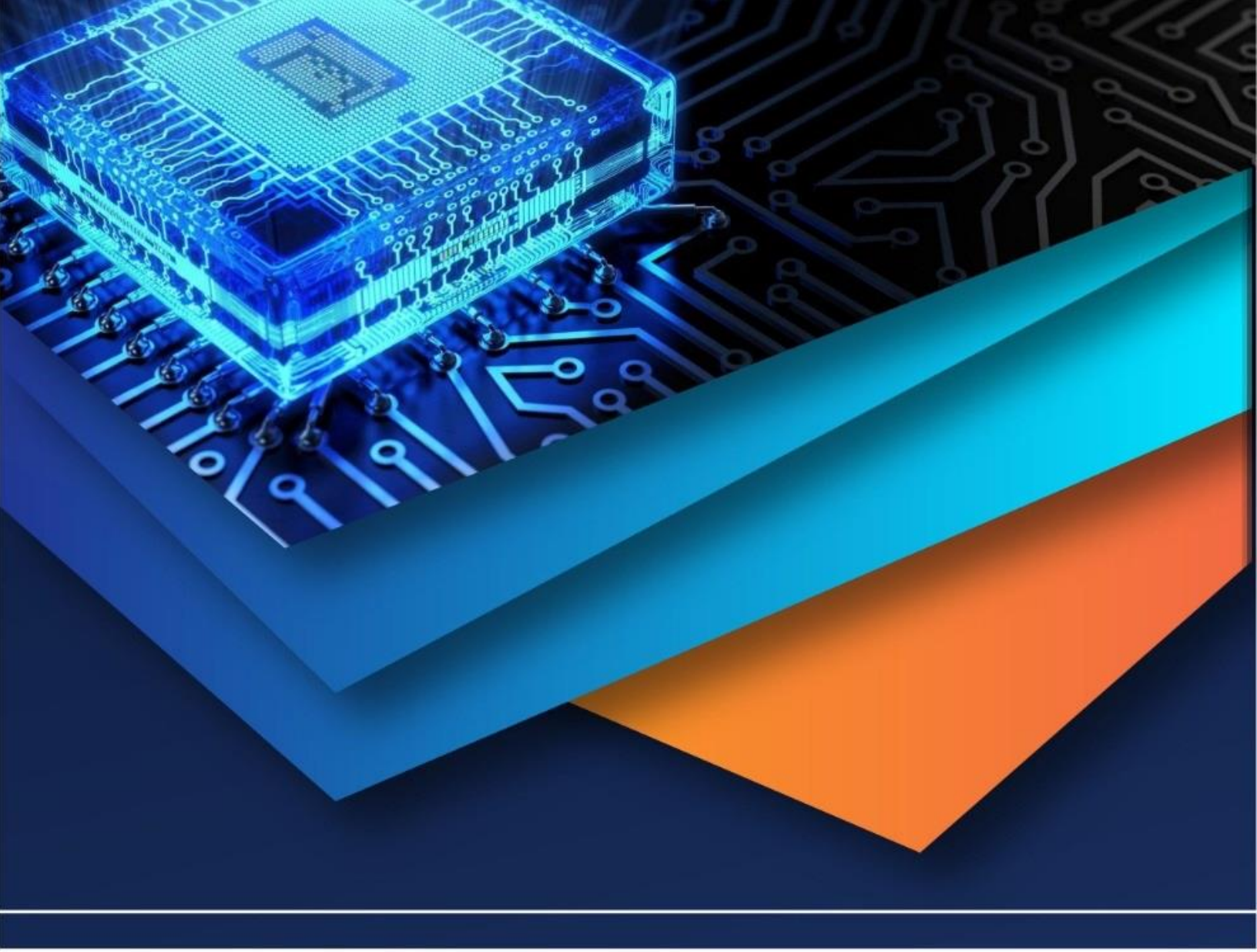
[8] Andre Brinkmann, Sascha Effert, "Snapshots and Continuous Data Replication in Cluster Storage Environments," Fourth International Workshop on Storage Network Architecture and Parallel I/O, IEEE,2008.

[9] Q. He, Z. Li, X. Zhang, "Data deduplication techniques," Future Information Technology and Management Engineering (FITME), vol. I, pp. 430-433, 2010.
[10] Maddodi.S, Attigeri G.V, Karunakar. A.K, "Data Deduplication Techniques and Analysis," Emerging Trends in Engineering and Technology (lCETET), pp 664 - 668, IEEE, 2010.
[11] Arasu, A., Ganti, V., Kaushik, R.: Efficient exact set-similarity joins. In: Proceedings of the 32nd International Conference on Very Large Data Bases (2006)
[12] Bilenko, M., Mooney, R.J.: On evaluation and training-set construction for duplicate detection. In: Proceedings of the KDD 2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation (2003).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)